

Quantifying Uncertainty in Transdimensional Markov Chain Monte Carlo Using Discrete Markov Models

Daniel W. Heck^{*1}, Antony M. Overstall², Quentin F. Gronau³, and
Eric-Jan Wagenmakers³

¹*Department of Psychology, University of Mannheim*

²*Mathematical Sciences, University of Southampton*

³*Department of Psychology, University of Amsterdam*

March 31, 2017

Abstract

Bayesian analysis often concerns an evaluation of models with different dimensionality as is necessary in, for example, model selection or mixture models. To facilitate this evaluation, transdimensional Markov chain Monte Carlo (MCMC) has proven to be a valuable tool. For instance, in case of model selection, this method relies on sampling a discrete model-indicator variable to estimate the posterior model probabilities. However, little attention has been paid to the precision of these estimates. If

^{*}Daniel W. Heck is PhD Student at the Graduate School for Economic and Social Sciences, University of Mannheim, Germany (e-mail: heck@uni-mannheim.de); Antony Overstall is Associate Professor in Statistics at the University of Southampton, United Kingdom (e-mail: A.M.Overstall@soton.ac.uk); Quentin F. Gronau is PhD Student at the Department of Psychology, University of Amsterdam, The Netherlands (e-mail: quentingronau@web.de); and Eric-Jan Wagenmakers is Full Professor at the Methodology Unit of the Department of Psychology, University of Amsterdam, The Netherlands (e-mail: ej.wagenmakers@gmail.com).

R code for all simulations is available at the Open Science Framework at <https://osf.io/kjrhz>, and the R package `MCMCprecision` is available at <https://github.com/danheck/MCMCprecision>.

only few switches occur between the models in the transdimensional MCMC output, precision may be low and assessment based on the assumption of independent samples misleading. Here, we propose a new method to estimate the precision based on the observed transition matrix of the indicator variable. Essentially, the method samples from the posterior of the stationary distribution, thereby assessing the uncertainty in the estimated posterior model probabilities. Moreover, the method provides an estimate for the effective sample size of the MCMC output. In two model-selection examples, we show that the proposed approach provides a good assessment of the uncertainty associated with the estimated posterior model probabilities.

Keywords: reversible jump MCMC, product space MCMC, convergence diagnostic, Bayesian model selection, posterior model probabilities, Bayes factor.

1. INTRODUCTION

Transdimensional Markov chain Monte Carlo (MCMC) methods provide an indispensable tool for the Bayesian analysis of models with varying dimensionality (Sisson, 2005). An important application is Bayesian model selection, where the aim is to estimate posterior model probabilities $p(\mathcal{M}_i \mid \mathbf{x})$ for a set of models \mathcal{M}_i , $i = 1, \dots, I$ given the data \mathbf{x} (Kass and Raftery, 1995). In order to ensure that the Markov chain converges to the correct stationary distribution, transdimensional MCMC methods such as reversible jump MCMC (Green, 1995) or the product space approach (Carlin and Chib, 1995) match the dimensionality of parameter spaces across different models (e.g., by adding parameters and link functions). Transdimensional MCMC methods have proven to be very useful for the analysis of many statistical models including capture-recapture models (Arnold, Hayakawa, and Yip, 2010), generalized linear models (Forster, Gill, and Overstall, 2012), factor models (Lopes and West, 2004), and mixtures models (Frühwirth-Schnatter, 2001), and are widely used in substantive applications such as selection of phylogenetic trees (Opgein-Rhein, Fahrmeir, and Strimmer, 2005), gravitational wave detection in physics (Karnesis, 2014), or cognitive models in psychology (Lodewyckx et al., 2011).

Crucially, transdimensional MCMC methods always include a discrete parameter Z with values in $1, \dots, I$ indexing the competing models. At iteration $t = 1, \dots, T$, posterior samples are obtained for the indicator variable $z^{(t)}$ and the model parameters, which are usually continuous and differ in dimensionality (for a review, see Sisson, 2005). For instance, a Gibbs sampling scheme can be adopted (Barker and Link, 2013), in which the indicator variable Z and the continuous model parameters are updated in alternating order. Such a sampler switches between models depending on the current values of the continuous parameters, and then updates these parameters in light of the current model \mathcal{M}_i conditionally on the value of $z^{(t)} = i$ (Barker and Link, 2013). Given convergence of the MCMC chain, the sequence $z^{(t)}$ follows a stationary distribution $\boldsymbol{\pi} = (\pi_1, \dots, \pi_I)$. Due to

the design of the sampler, this distribution is identical to the posterior model probabilities of interest, $\pi_i = p(\mathcal{M}_i \mid \mathbf{x})$ and, given uniform model priors $p(\mathcal{M}_i) = 1/I$, also proportional to the marginal likelihoods $p(\mathbf{x} \mid \mathcal{M}_i)$. Hence, transdimensional MCMC samplers can be used to directly estimate these posterior probabilities as the relative frequencies of samples $z^{(t)}$ falling into the I categories, $\hat{\pi}_i = 1/T \sum_t \mathbb{I}(z^{(t)} = i)$, where \mathbb{I} is the indicator function. Due to the ergodicity of the Markov chain, this estimator is ensured to be asymptotically unbiased (Green, 1995; Carlin and Chib, 1995).

Usually, dependencies due to MCMC sampling are taken into account for continuous parameters (Cowles and Carlin, 1996). In contrast, however, the estimate $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \dots, \hat{\pi}_I)$ based on the sequence of discrete samples $z^{(t)}$ is usually reported without quantifying estimation uncertainty. Often, the samples $z^{(t)}$ are correlated to a substantial, but unknown degree because of infrequent switching between models. This is illustrated in Figure 1, which shows a sequence of independent and correlated samples $z^{(t)}$ in Panels A and B, respectively. Inference about the stationary distribution $\boldsymbol{\pi}$ should be more reliable in the first case compared to the second case, in which the autocorrelation reduces the amount of information available about $\boldsymbol{\pi}$. Moreover, the standard error $\text{SE}(\hat{\pi}_i) = \sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)/T}$ that assumes independent sampling will obviously underestimate the true variability of the estimate $\hat{\boldsymbol{\pi}}$ (Green, 1995; Sisson, 2005). To obtain a measure of precision, Green (1995) proposed running several independent MCMC chains $c = 1, \dots, C$ and computing the standard error of the estimate $\hat{\boldsymbol{\pi}}^{(c)}$ across these independent replications. However, for complex models, this method might require a substantial amount of additional computing time for burn-in and adaption and thus can be infeasible in practice.

Assessing the precision of the estimate $\hat{\boldsymbol{\pi}}$, which depends on the autocorrelation of the sequence of discrete samples $z^{(t)}$, is of major importance. In case of model selection, it must be ensured that the estimated posterior probabilities $p(\mathcal{M}_i \mid \mathbf{x})$ are sufficiently precise for drawing substantive conclusions. This issue is especially important when estimating a ratio of marginal probabilities, that is, the Bayes factor $B_{ij} = p(\mathbf{x} \mid \mathcal{M}_i)/p(\mathbf{x} \mid \mathcal{M}_j)$ (Jeffreys,

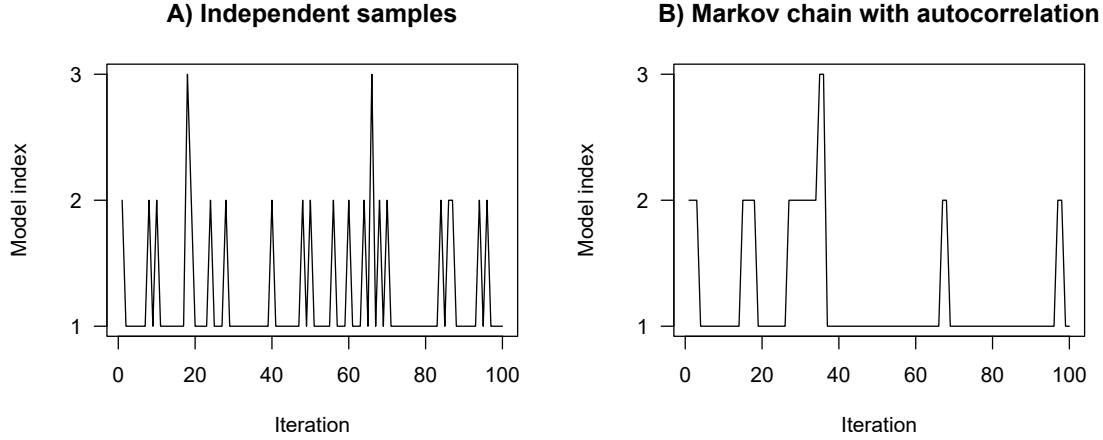


Figure 1: Illustration of the MCMC iterations of the model-index parameter $z^{(t)}$. Whereas samples are independent multinomial samples in Panel A, samples are drawn from a Markov chain with positive autocorrelation in Panel B.

1961). More generally, it is often of interest to compute the effective sample size, that is, the number of independent samples that would provide the same amount of information as the given MCMC output. Besides providing an intuitive measure of precision, a minimum effective sample size can serve as a principled and theoretically justified stopping rule for MCMC sampling (Gong and Flegal, 2016). Software for convergence diagnostics such as the R package `coda` (Plummer et al., 2006) estimate the effective sample size for continuous parameters by computing the spectral density at zero (Heidelberger and Welch, 1981). However, for transdimensional MCMC methods, this approach neglects the discreteness of $z^{(t)}$. Depending on the numerical labels for the different models (e.g., using the labels $(1, 2, 3, 4)$ vs. $(1, 4, 2, 3)$), the spectral-density estimation can lead to widely varying and arbitrary results (see Section 4).

In summary, transdimensional MCMC is a very important and popular method for Bayesian inference (Sisson, 2005). However, little attention has been paid to the analysis of the resulting MCMC output, which requires that one takes into account the autocorrelation as well as the discrete nature of the model indicator variable. As a solution, we propose to fit a discrete Markov model to the MCMC output $z^{(t)}$ to assess the precision of

the estimated stationary distribution $\hat{\boldsymbol{\pi}}$. Note that our approach differs from diagnostics previously proposed to monitor convergence of transdimensional MCMC methods. These diagnostics usually compare the variance of $z^{(t)}$ and other continuous parameters across and within chains and models (Brooks and Giudici, 2000; Castelloe and Zimmerman, 2002), similar to the widely used potential scale reduction factor (Gelman and Rubin, 1992). Other methods estimate the convergence rate of transdimensional MCMC chains, rely on Kolmogorov-Smirnov or χ^2 tests (Brooks, Giudici, and Philippe, 2003), or use distance-based diagnostics (Sisson and Fan, 2007). However, none of these methods estimates the precision of the point estimate $\hat{\boldsymbol{\pi}}$.

2. METHOD

2.1. Posterior Distribution of Model Probabilities

To estimate the fixed, but unknown stationary distribution $\boldsymbol{\pi}$ (e.g., the posterior model probabilities), we explicitly model the sequence of observed samples as a random variable $Z^{(t)}$ ($t = 1, \dots, T$) by assuming that it has emerged from a discrete, homogeneous Markov chain $\mathcal{M}^{\text{Markov}}$. For this purpose, we only consider the marginal distribution of the discrete indicator variable and define the transition matrix \boldsymbol{P} with the switching probabilities $p_{ij} = P(Z^{(t+1)} = j \mid Z^{(t)} = i)$ for all $i, j = 1, \dots, I$. Accordingly, \boldsymbol{P} is a probability matrix with non-negative entries and rows summing to one, i.e. $\sum_{j=1}^I p_{ij} = 1$. For this Markov model, the resulting probability distribution at iteration t is given by multiplying the transposed initial distribution $\boldsymbol{\pi}'_0$ by the transition matrix t times, $P(Z^{(t)} = i) = [\boldsymbol{\pi}'_0 \boldsymbol{P}^t]_i$. Within the model $\mathcal{M}^{\text{Markov}}$, the transition matrix \boldsymbol{P} is a free parameter that is to be estimated based on the sampled sequence $z^{(t)}$.

Due to the construction of the transdimensional MCMC sampler, $Z^{(t)}$ has $\boldsymbol{\pi}$ as a stationary distribution. Therefore, the transition matrix \boldsymbol{P} in the Markov model $\mathcal{M}^{\text{Markov}}$

satisfies the condition

$$\boldsymbol{\pi}'\mathbf{P} = \mathbf{1} \cdot \boldsymbol{\pi}', \quad (1)$$

which implies that the probability vector $\boldsymbol{\pi}$ is the left eigenvector (normalized to sum to one; Anderson and Goodman, 1957) of the matrix \mathbf{P} with eigenvalue one. This eigenvector exists if the Markov chain is finite and irreducible (Stewart, 2009, Ch. 9). Hence, given an estimate for \mathbf{P} , we can directly compute $\boldsymbol{\pi}$ as the corresponding eigenvector.

However, we are less interested in a point estimate $\hat{\boldsymbol{\pi}}$ of the stationary distribution but rather in the precision of this estimate. For this purpose, the sequence $z^{(t)}$ is summarized by the sufficient statistic \mathbf{N} (Anderson and Goodman, 1957), the observed matrix of transition frequencies, where n_{ij} is the number of switches from $z^{(t)} = i$ to $z^{(t+1)} = j$. Conditional on the MCMC output, the posterior distribution of \mathbf{P} can be approximated by drawing $r = 1, \dots, R$ samples according to

$$\mathbf{P}^{(r)} \sim p(\mathbf{P} \mid \mathbf{N}, \mathcal{M}^{\text{Markov}}). \quad (2)$$

By computing the (normalized) eigenvectors with eigenvalue one (Eq. 1), posterior samples of the stationary distribution $\boldsymbol{\pi}$ are directly obtained. As a prior for the transition matrix \mathbf{P} , we assume independent Dirichlet distributions with parameter $\epsilon \geq 0$ for each of the rows,

$$\mathbf{p}_i \equiv (p_{i1}, \dots, p_{iI}) \sim \mathcal{D}(\epsilon, \dots, \epsilon). \quad (3)$$

Therefore, independent posterior samples $\mathbf{P}^{(r)}$ can efficiently be drawn from the conjugate posterior distribution,

$$\mathbf{p}_i^{(r)} \sim \mathcal{D}(n_{i1} + \epsilon, \dots, n_{iI} + \epsilon). \quad (4)$$

With regard to the prior parameter ϵ , small values should be chosen to reduce its influence on the estimation of \mathbf{P} . Note that this choice becomes less influential as the number of MCMC samples increases (especially if the row sums of \mathbf{N} are large). Here, we

use the prior $\epsilon = 1/I$, which has an impact equivalent to one observation for each row of the observed transition matrix \mathbf{N} . This prior has proved to be numerically robust in the two examples in Sections 4 and 5, and resulted in point estimates close to the default i.i.d. estimates.

Alternatively, the improper prior $\epsilon = 0$ can be used, which minimizes the impact of the prior on the estimated stationary distribution. Note that this prior also ensures that the results do not hinge on the set of models that could possibly be sampled, but were never actually observed in the sequence $z^{(t)}$. For such unsampled models, the corresponding rows and columns of the observed transition matrix \mathbf{N} are filled with zeros. With $\epsilon = 0$, the relevant eigenvector of the posterior $\mathbf{P} \mid \mathbf{N}$ is thus identical to that of a reduced matrix $\tilde{\mathbf{P}} \mid \tilde{\mathbf{N}}$ that includes only the transitions for the subset of models sampled in $z^{(t)}$. Moreover, if the set of competing models is large, \mathbf{N} is likely to be a sparse matrix because many transitions will never occur. When choosing $\epsilon = 0$, $\mathbf{P}^{(r)}$ will also be a sparse matrix when sampling from the conjugate Dirichlet distribution $\mathcal{D}(n_{i1}, \dots, n_{iI})$, which facilitates an efficient computation of the eigenvectors $\boldsymbol{\pi}^{(r)}$. However, in our simulations, this improper Dirichlet prior proved to be numerically unstable and resulted in more variable point estimates than the proper prior $\epsilon = 1/I$.

As a third alternative, the prior can be adapted to the structure of specific transdimensional MCMC implementations, which only implement switches to a small subset of the competing models. For instance, in variable selection, regression parameters are often added or removed one at a time, resulting in a birth-death process (Stephens, 2000). For these kinds of samplers, the Dirichlet parameters ϵ_{ij} can be set to zero selectively. However, such adjustments will be dependent on the chosen MCMC sampling scheme. Therefore, we propose the weakly informative prior $\epsilon = 1/I$ as a default, which provides a good compromise of being very general and numerically robust, while having a small effect on the posterior.

2.2. Precision

Based on the posterior samples $\boldsymbol{\pi}^{(r)}$, it is straightforward to estimate the stationary distribution by the posterior mean $\hat{\boldsymbol{\pi}}$ (alternatively, the median or mode may be used). More importantly, however, estimation uncertainty due to the transdimensional MCMC method can directly be assessed by plotting the estimated posterior densities for each π_i . To quantify the precision of the estimate $\hat{\boldsymbol{\pi}}$, one can report posterior standard deviations or credibility intervals for the components $\hat{\pi}_i$. Note that these component-wise summary statistics are most useful if the number of models I is relatively small.

For very large numbers of sampled models, the assessment of estimation uncertainty can be focused on the subset of models with the highest posterior model probabilities. Besides summarizing the estimated posterior model probabilities, estimation uncertainty for the k best-performing models can also be assessed by computing ranks for each of the posterior samples $\boldsymbol{\pi}^{(r)}$. Then, the variability of these model rankings across the R samples can be summarized, for instance, by the percentage of identical rank orders for the k best-performing models, or the percentages of how often each model is included within the subset of the k best-performing models (i.e., has a rank smaller or equal to k).

In case of model selection, dispersion statistics such as the posterior standard deviation are also of interest with respect to the Bayes factor B_{ij} (Kass and Raftery, 1995). To judge the estimation uncertainty for the Bayes factor, one can evaluate the corresponding posterior distribution by computing the derived quantities $B_{ij}^{(r)} = \pi_i^{(r)} / \pi_j^{(r)}$ (given uniform prior model probabilities). Precision can also be assessed for model-averaging contexts when comparing subsets of models against each other (e.g., regression models including a specific effect vs. those not including it). Given such disjoint sets of model indices $M_s \subset \{1, \dots, I\}$, the posterior probability for each subset of models is directly obtained by summing the posterior samples $\pi_i^{(r)}$ for all $i \in M_s$. Note that it is invalid to aggregate across model subsets before applying the proposed Markov approach because functions of

discrete Markov chains (e.g., collapsing the I original states into a subset of S states) are not Markovian in general (Burke and Rosenblatt, 1958).

2.3. Effective Sample Size

Besides quantifying estimation uncertainty, the posterior samples $\boldsymbol{\pi}^{(r)}$ can be used to compute the effective sample size for the transdimensional MCMC output. For this purpose, we consider the benchmark model \mathcal{M}^{iid} under the ideal scenario of drawing independent samples $\tilde{z}^{(t)}$ from the categorical distribution with probabilities $\tilde{\boldsymbol{\pi}}$ (which is equivalent to sampling from a multinomial distribution). For this model, we also assume a Dirichlet prior, but this time directly on the stationary distribution, $\tilde{\boldsymbol{\pi}} \sim \mathcal{D}(\gamma, \dots, \gamma)$ with a fixed parameter $\gamma \geq 0$. Since the prior is conjugate, the posterior for the estimated distribution is given by

$$\tilde{\boldsymbol{\pi}} \mid \tilde{\mathbf{n}} \sim \mathcal{D}(\tilde{n}_1 + \gamma, \dots, \tilde{n}_I + \gamma), \quad (5)$$

based on the observed frequencies $\tilde{n}_i = \mathbb{I}(z^{(t)} = i) = \sum_j n_{ij}$. By considering only the proportion that each model is visited, the transition frequencies are rendered irrelevant in this i.i.d. approach, thereby ignoring possible dependencies in the sample sequence $z^{(t)}$.

Given the output of a transdimensional MCMC chain, we can now compare the empirical posterior of $\boldsymbol{\pi}$ derived from the model $\mathcal{M}^{\text{Markov}}$ against the theoretically expected posterior $\tilde{\boldsymbol{\pi}}$ under the model \mathcal{M}^{iid} to estimate the sample size $\tilde{T} = \sum_i \tilde{n}_i$. For this purpose, a Dirichlet distribution with parameters $\alpha_1, \dots, \alpha_I$ is fitted to the samples $\boldsymbol{\pi}^{(r)}$, which can be accomplished by an efficient fixed-point iteration scheme (Minka, 2000). Second, a comparison of the estimated Dirichlet parameters with the conjugate posterior in Eq. 5 yields $\hat{\alpha}_i \approx \tilde{n}_i + \gamma$. Therefore, after subtracting the prior sample size $I^2\epsilon$ of the $I \times I$ transition matrix \mathbf{P} (Eq. 3), the effective sample size under the assumption of independent

sampling from a multinomial distribution can be estimated as

$$\hat{T}_{\text{eff}} = \sum_{i=1}^I \hat{\alpha}_i - I^2 \epsilon \approx \sum_{i=1}^I \tilde{n}_i + I\gamma. \quad (6)$$

To minimize the impact of the prior, one can assume the improper prior $\gamma = 0$ as a default. Importantly, this estimate takes the discreteness of the indicator variable Z into account and does not change under permutations of arbitrary, numerical values of the model indices.

2.4. Remarks

If output from multiple independent chains $c = 1, \dots, C$ is available, the transition frequency matrices $\mathbf{N}^{(1)}, \dots, \mathbf{N}^{(C)}$ can simply be summed before applying the method. This follows directly from Bayesian updating of the stationary distribution $\boldsymbol{\pi}$. Essentially, each chain provides independent evidence for the posterior, which is reflected by using the sums $\sum_c n_{ij}^{(c)}$ for the conjugate Dirichlet prior in Eq. 4. Note that this feature can be used to compare the efficiency of many short versus few long MCMC chains.

The proposed method bears a resemblance to the convergence diagnostic of Brooks, Giudici, and Philippe (2003). To assess the convergence rate of a transdimensional MCMC chain, Brooks, Giudici, and Philippe (2003) computed the second largest eigenvalue of the maximum-likelihood estimate of the transition matrix, $\hat{p}_{ij} = n_{ij} / \sum_k n_{ik}$. Since $\hat{\mathbf{P}}$ is a probability matrix, this eigenvalue must be smaller than one and provides a measure of the dependency of the samples $z^{(t)}$.

Note that the simplifying assumptions underlying our approach are not guaranteed to hold. Whereas samples of the full model space $(z^{(t)}, \boldsymbol{\theta}^{(t)})$ necessarily follow a Markov process by construction, this does not imply that the samples $z^{(t)}$ follow a Markov chain marginally (Brooks, Giudici, and Roberts, 2003; Lodewyckx et al., 2011). Intuitively, this is due to the fact that the transition probabilities depend on the exact locations of the MCMC sampler in each of the models' parameter spaces. However, in Sections 4 and 5 we show in two

empirical examples that the proposed simplification (i.e., fitting a Markov chain of order one) is sufficient to account for autocorrelations in the samples $z^{(t)}$ in practice.

The proposed method can be applied irrespective of specific transdimensional MCMC implementations and requires only the sampled sequence $z^{(t)}$ of the discrete parameter or the matrix \mathbf{N} with the observed frequency of transitions. In the R package `MCMCprecision` (Heck et al., 2017), we provide an implementation that relies on the efficient computation of eigenvectors in the C++ library `Armadillo` (Sanderson and Curtin, 2016), accessible in R via the package `RcppArmadillo` (Eddelbüttel and Sanderson, 2014). On a notebook with an Intel® i5-3320M processing unit, drawing $R = 1,000$ samples from the posterior distribution for 10 (100) sampled models requires approximately 70 milliseconds (10 seconds).

3. ILLUSTRATION: EFFECT OF AUTOCORRELATION

Before applying the proposed method to actual MCMC output, we first illustrate its use in an idealized setting. To investigate the effect of autocorrelation in the case of discrete parameters, we generate sequences $z^{(t)}$ of length $T = 1,000$ from the Markov model $\mathcal{M}^{\text{Markov}}$ for a given stationary distribution $\boldsymbol{\pi} = (.85, .13, .02)'$. To induce autocorrelation, we define a mixture process for each iteration t . With probability β , the discrete indicator variable will be identical to the current model, $z_{t+1} = z_t$. In contrast, with probability $1 - \beta$, the value z_{t+1} is sampled from the given stationary distribution $\boldsymbol{\pi}$. Thereby, increasing values of β result in larger autocorrelation of the sequence $z^{(t)}$.

For varying levels of $\beta = 0, 0.1, \dots, 0.8$, we sampled 500 replications, applied the proposed method and computed the precision of the estimate $\hat{\boldsymbol{\pi}}^{\text{Markov}}$. Besides the posterior SD, we were interested in the coverage probability of the data-generating value $\boldsymbol{\pi}$ being in the 90% credibility interval. As a benchmark, we also computed the corresponding posterior SD under the (false) assumption that the samples $z^{(t)}$ were independently drawn by fitting the model \mathcal{M}^{iid} with the Dirichlet prior parameter $\gamma = 0$.

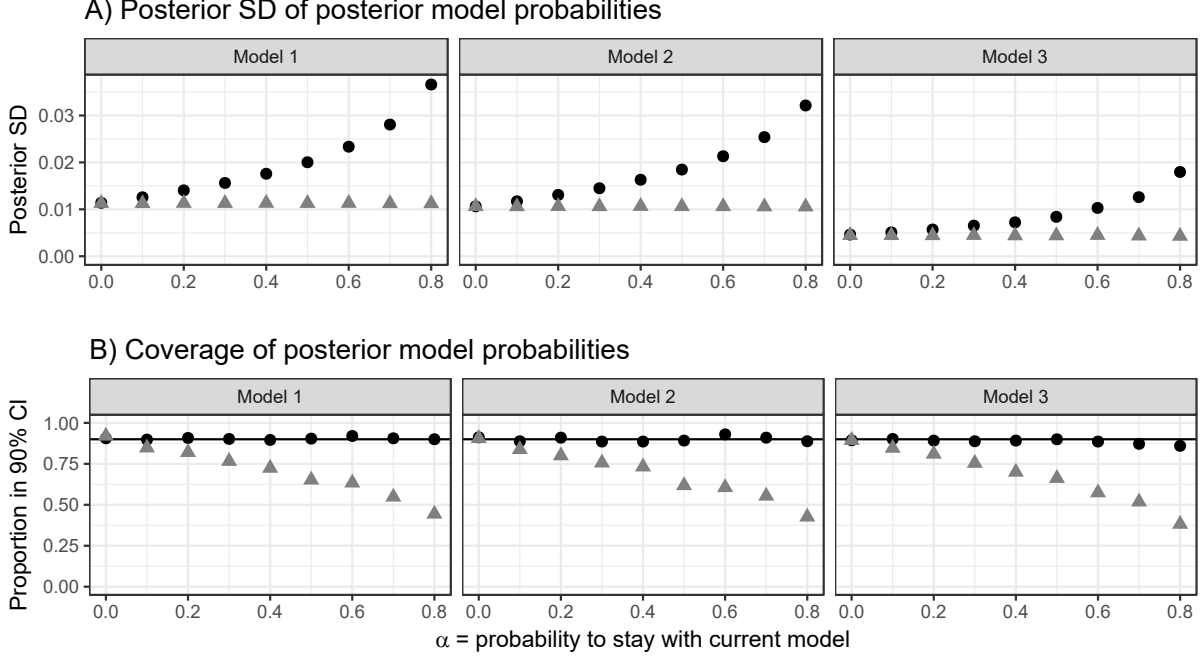


Figure 2: Estimation uncertainty for the stationary distribution π . (A) The Markov-method (black dots) correctly indicates that estimation error of the posterior model probabilities increases as autocorrelation increases. When assuming i.i.d. samples (gray triangles), the estimated precision does not depend on the autocorrelation. (B) Proportion of 500 replications for which the 90% CI intervals include the data-generating stationary distribution π .

Figure 2 shows the result of this simulation. In Panel A, the estimation uncertainty increases for larger values of β , thereby taking the increasing autocorrelation into account. In contrast, the model \mathcal{M}^{iid} does assume independence a priori. Thereby, the posterior uncertainty is independent of β . As a result of this, the 90% credibility interval is less likely to include the data-generating value π as shown in Panel B, whereas the Markov model provides an accurate description of the estimation uncertainty.

4. VARIABLE SELECTION IN LOGISTIC REGRESSION

In the following, we apply the proposed method to the problem of selecting variables in a logistic regression, an example introduced by Dellaportas, Forster, and Ntzoufras (2000) to highlight the implementation of transdimensional MCMC in BUGS (see also Dellaportas,

Table 1: Logistic regression data set by Healy (1988).

Condition (A)	Antitoxin (B)	Death	Survival
More Severe	Yes	15	6
	No	22	4
Less Severe	Yes	5	15
	No	7	5

Forster, and Ntzoufras, 2002; Ntzoufras, 2002). Table 1 shows the frequencies of deaths and survivals conditional on severity and whether patients received treatment (i.e., antitoxin medication; Healy, 1988). To emphasize the importance of considering estimation uncertainty for the posterior model probabilities, we compare the efficiency of two transdimensional MCMC approaches, which can both be implemented in JAGS (Plummer, 2003).

The full logistic regression model assumes a Bernoulli distribution \mathcal{B} of the survival frequencies x_{jl} and a linear model on the logit-transformed survival probabilities p_{jl} ,

$$x_{jl} \sim \mathcal{B}(p_{jl}, n_{jl}) \quad (7)$$

$$\log \left(\frac{p_{jl}}{1 - p_{jl}} \right) = \beta_0 + \beta_1 a_j + \beta_2 b_l + \beta_3 (ab)_{jl}, \quad j, l = 1, 2 \quad (8)$$

where n_{jl} are the total number of patients in condition jl and β the regression coefficient for the effect-coded variables a_j , b_l , and $(ab)_{jl}$. Variable selection is required to choose between $I = 5$ models: the intercept-only model, the three main effect models A, B, and A+B, and the model AB that includes the interaction. For comparability, we use the same priors as Dellaportas, Forster, and Ntzoufras (2000) and assume a centered Gaussian prior with variance $\sigma^2 = 8$ for each regression parameter, $\beta_k \sim \mathcal{N}(0, 8)$. Moreover, the model probabilities were set to be uniform, $p(\mathcal{M}_i) = 1/5$. Note that efficiency can be increased by selecting prior probabilities that result in approximately uniform posterior probabilities $p(\mathcal{M}_i | \mathbf{x}) \approx 1/I$ (Lodewyckx et al., 2011). Moreover, nonuniform prior probabilities might

be used to protect against multiple testing issues (i.e., Bayes multiplicity; Scott and Berger, 2010).

One of the two implemented transdimensional MCMC approaches uses unconditional priors (Kuo and Mallick, 1998, KM98) and includes indicator variables $\gamma_{ik} \in \{0, 1\}$ for each regression coefficient β_k in model \mathcal{M}_i . The parameter γ_i determines which regression coefficients are included by removing some of the additive terms of the linear model in Equation 8. Details about the full and conditional posterior distributions are provided by Dellaportas, Forster, and Ntzoufras (2000, p. 7).

As a second transdimensional MCMC approach, we implemented the method of Carlin and Chib (1995; CC95), which stacks up all model parameters into a new parameter $\boldsymbol{\theta} = (z, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_I)$, where $\boldsymbol{\beta}_i$ is the vector of regression parameters of model \mathcal{M}_i . Thereby, this approach samples a total of 12 regression parameters along with the indicator variable z . Note that the method of Carlin and Chib (1995) uses pseudo-priors $p(\boldsymbol{\beta}_i | \mathcal{M}_{i'})$, $i \neq i'$, that do not influence the statistical inference about $p(\mathbf{x} | \mathcal{M}_i)$ and $p(\boldsymbol{\beta}_i | \mathbf{x}, \mathcal{M}_i)$. However, these pseudo-priors determine the conditional proposal probabilities $p(z | \mathbf{x}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_I)$ of switching between the models and thereby affect the efficiency of the MCMC chain. In substantive applications, these pseudo-priors should be chosen to match the posterior $p(\boldsymbol{\beta}_i | \mathcal{M}_i)$ in order to ensure high probabilities of switching between the models (cf. Carlin and Chib, 1995; Barker and Link, 2013). Here, however, we did not optimize the sampling scheme and use $\boldsymbol{\beta}_{ik} | \mathcal{M}_{i'} \sim \mathcal{N}(0, 8)$ for the pseudo-priors to illustrate that our method can correctly detect the lower precision resulting from this suboptimal choice.

Figure 3 shows the estimated posterior distribution of the posterior model probabilities using one Markov chain with 21,000 iterations (including 1,000 burn-in samples). The vertical black lines show the correct reference values, approximated by eight independent chains with one million samples each. As expected, the assumption that $z^{(t)}$ are sampled independently results in overconfidence in the point estimates of the CC95 approach. For all models, the corresponding posterior distributions miss the correct value and do not

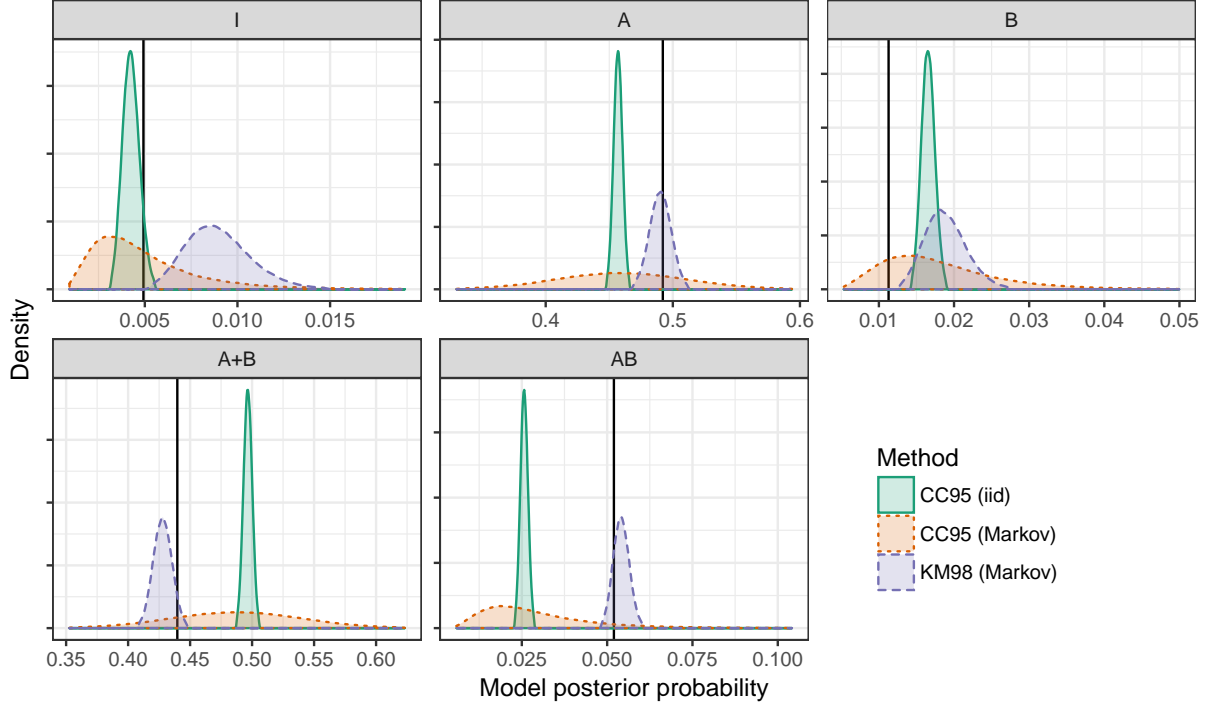


Figure 3: Posterior distribution of the posterior model probabilities π in the logistic regression example based on the Markov and the i.i.d. model. Vertical black lines show the target values (CC95 = Carlin and Chib, 1995; KM98 = Kuo and Mallick, 1998).

identify this estimation uncertainty. In contrast, the proposed Markov approach results in a posterior distribution that covers the target values with sufficiently high probability. Moreover, the novel estimation method reveals that the KM98 implementation has a higher precision compared to the (intentionally not optimized) CC95 approach.

To test the validity of the proposed method more rigorously, we replicated the previous analysis 100 times. Thereby, the estimated precision can be compared against the actual sampling variability of the estimated model probabilities. For both transdimensional MCMC methods, Table 2 shows the mean estimated model probabilities in percent. Across replications, the point estimates from the Markov and the i.i.d. approach were similar with mean absolute differences smaller than 0.02% and 0.49% for the KM98 and CC95 implementations, respectively. To judge whether the estimated precision (i.e., the mean posterior standard deviations \overline{SD}_i and \overline{SD}_M) is valid, Table 2 shows the empirical SD of the estimates

Table 2: Estimated posterior model probabilities in percent.

Model	Kuo and Mallick (1998)				Carlin and Chib (1995)			
	Mean($\hat{\pi}$)	\overline{SD}_i	\overline{SD}_M	SD($\hat{\pi}$)	Mean($\hat{\pi}$)	\overline{SD}_i	\overline{SD}_M	SD($\hat{\pi}$)
1	0.46	0.05	0.11	0.14	0.55	0.05	0.24	0.15
A	49.12	0.35	0.86	0.97	48.98	0.35	4.93	5.44
B	1.06	0.07	0.17	0.29	1.08	0.07	0.43	0.24
A+B	44.10	0.35	0.78	0.87	43.56	0.35	5.12	4.97
AB	5.26	0.16	0.24	0.22	5.83	0.16	2.54	1.32

Note: Posterior model probability estimates $\hat{\pi}$ are shown in percent. Mean($\hat{\pi}$) and SD($\hat{\pi}$) were computed across 100 replications. As a measure for the estimated precision, means of the posterior SD are shown (\overline{SD}_i assumes independent sampling; \overline{SD}_M assumes a Markov chain model).

$\hat{\pi}^{\text{Markov}}$ across replications. The results clearly show that the assumption of independent samples $z^{(t)}$ leads to an overconfident assessment of the precision for the estimated model probabilities, whereas the proposed Markov approach provides a good estimate of the actual estimation uncertainty. Moreover, for the MCMC method by Carlin and Chib (1995), the larger SDs indicate a smaller efficiency compared to the unconditional prior approach by Kuo and Mallick (1998). This theoretically expected result is likely to be due to the suboptimal choice of pseudo-priors. However, note that this difference in efficiency may be overlooked when merely computing relative proportions based on the sampled indicator variable $z^{(t)}$ (i.e., when implicitly assuming independent samples).

The higher efficiency of the KM98 approach becomes even clearer when assessing the mean effective sample size across replications, which was estimated to be 4,514 compared to only 163 for the CC95 method. Note that commonly used estimators of effective sample size (e.g., Plummer et al., 2006) depend on the exact numerical labels of the model-indicator variable Z . To illustrate this, Figure 4 shows the estimated sample size for all 120 permutations of the indices $(1, \dots, 5)$ for a fixed sequence $z^{(t)}$, computed by the spectral decomposition available in the R package coda (Plummer et al., 2006). This estimate varied

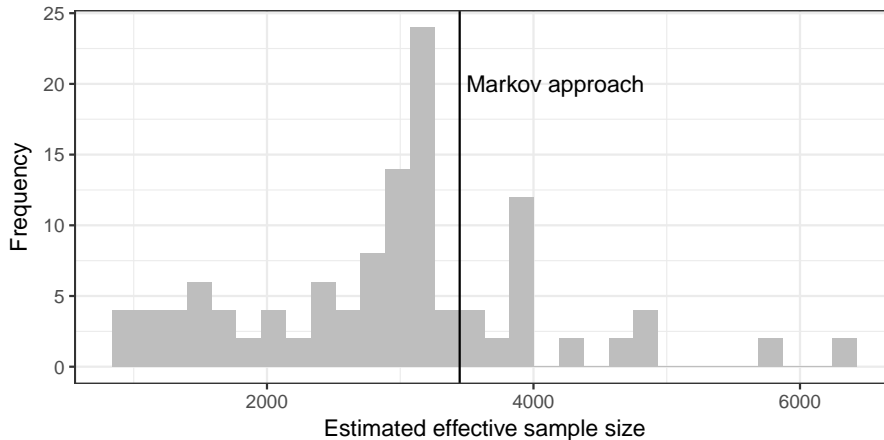


Figure 4: Effective sample size as estimated by the spectral density at zero (Plummer et al., 2006) for all permutations of the model indicator labels of the MCMC output $z^{(t)}$ (based on 20,000 samples of the method by Kuo and Mallick, 1998).

considerably depending on the arbitrary labeling of the models. In contrast, the proposed Markov approach results in a well-defined, invariant estimate by explicitly accounting for the discreteness of Z .

Finally, we show that the posterior samples $\pi^{(t)}$ of the model $\mathcal{M}^{\text{Markov}}$ can directly be used to assess the uncertainty of Bayes factor estimates. For instance, substantive applications could be interested in testing whether to include the interaction term of condition (A) and treatment (B) in a logistic regression model. Given the output of a single MCMC run with 20,000 samples, Figure 5 shows the resulting posterior distribution of the Bayes factor $B_{A+B,AB}$ in favor for the absence of an interaction. Similar to the posterior model probabilities, the i.i.d. approach results in overconfidence regarding the estimate and most of the probability mass excludes the correct value 8.50 (approximated with a precision of $\text{SD} = 0.014$). In contrast, the Markov approach corrects for dependencies in the samples $z^{(t)}$ and includes the correct value. The same pattern emerged across 100 replications, that is, the mean estimated SD of the Bayes factor matched the corresponding empirical SD of the Bayes factor estimates (KM98: 0.40 vs. 0.47; CC95: 6.93 vs. 6.24). When using trans-dimensional MCMC, Bayes factors cannot be expected to be reliably estimated if models

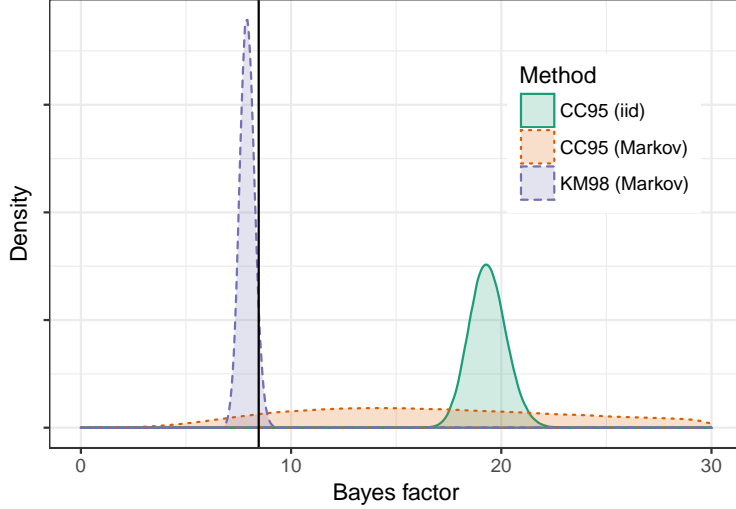


Figure 5: Posterior distribution for the Bayes factor in favor of Model A+B vs. AB. The vertical black line shows the target value (CC95 = Carlin and Chib, 1995; KM98 = Kuo and Mallick, 1998).

are never or very infrequently sampled (e.g., Model 1 in Table 2). For instance, the Bayes factor $B_{A,B} \approx 44.4$ was estimated very imprecisely even in the KM98 approach (mean SD = 7.5; empirical SD = 14.8). To obtain more precise Bayes factor estimates in the presence of infrequently sampled models, it is recommended to rerun the transdimensional MCMC chain including only the two relevant models of interest (Lodewyckx et al., 2011).

5. LOG-LINEAR MODELS FOR A 2^6 CONTINGENCY TABLE

The application of the proposed method is also feasible in realistic scenarios with hundreds of sampled models. To illustrate this, we reanalyzed the 2^6 complete contingency table by Edwards and Havránek (1985), which includes six risk factors for coronary heart disease (i.e., smoking, strenuous mental work, strenuous physical work, systolic blood pressure, ratio of α and β lipoproteins, and family anamnesis of coronary heart disease). We are interested in finding the most parsimonious log-linear model, which accounts for the cell frequencies y_j of cell j ($j = 1, \dots, 2^6$) by assuming a Poisson distribution with mean μ_j

and

$$\log \mu_j = \phi + \mathbf{x}'_j \boldsymbol{\beta}, \quad (9)$$

where ϕ is the intercept, $\boldsymbol{\beta}$ the vector of regression parameters, and \mathbf{x}'_j the (transposed) design vector, which selects the elements of $\boldsymbol{\beta}$ included for modeling cell j . Here, we consider the class of hierarchical log-linear models that only allow the inclusion of an interaction if the corresponding marginal effects and lower interaction terms are included in the model as well (e.g., Overstall and King, 2014a).

To select between all 7.8 million possible hierarchical log-linear models (Dellaportas and Forster, 1999), we use the reversible jump algorithm proposed by Forster, Gill, and Overstall (2012), which is implemented in the R package `conting` (Overstall and King, 2014b). Assuming a unit information prior (Ntzoufras, Dellaportas, and Forster, 2003), we sampled 100,000 iterations, discarded 10,000 as burn-in, and applied the proposed Markov chain method by drawing 2,000 samples for the posterior model probabilities. To assess whether the estimated uncertainty accurately quantifies sampling variability, we ran 100 replications initialized with randomly chosen models.

Across replications, 4,484 models were sampled (on average, 567.6 per replication). Table 3 shows the 10 models with the highest posterior probabilities. The relatively large posterior standard deviations of the estimated posterior model probabilities indicate that the samples $z^{(t)}$ are autocorrelated to a substantial degree, despite the large number of iterations. This is also reflected by the effective sample size, which was estimated to be $\hat{T}_{\text{eff}} = 4,484$ on average (SD = 186), approximately 5% of the number of iterations after burn-in.

Table 3 also shows means and standard deviations of the sampled rank R for the models with the highest posterior probability, indicating that estimation uncertainty (i.e., the posterior SD) increased for models with smaller posterior probabilities. Moreover, the proportion of posterior samples is shown for which the sampled rank R was identical to the

rank across all replications ($R = \#$) and smaller than or equal to 10 ($R \leq 10$). According to these proportions, exact ranks were estimated precisely only for the two best models, whereas the set of the 10 models with highest posterior probabilities was relatively stable across posterior samples (with the exception of model 10). Importantly, the mean estimated probabilities $\overline{P(R = \#)}$ and $\overline{P(R \leq 10)}$ matched the corresponding empirical proportions across replications.

Note that these results regarding estimation uncertainty are in line with our expectations — if models have small posterior probabilities, they are also sampled infrequently, which in turn results in estimation uncertainty. To quantify this variability, the proposed Markov chain approach provides an estimate for the achieved precision to assess the quality of the results and to find an appropriate stopping rule for MCMC sampling.

Table 3: Models with the highest posterior probability for the 2^6 contingency table.

#	Model	Posterior model probabilities $\boldsymbol{\pi}$				Rank R						
		Mean($\hat{\boldsymbol{\pi}}$)	\overline{SD}_i	\overline{SD}_M	SD($\hat{\boldsymbol{\pi}}$)	Mean(R)	$\overline{SD}(\overline{R})$	SD(\overline{R})	$\overline{P(R = \#)}$	$R = \#$	$\overline{P(R \leq 10)}$	$R \leq 10$
1	CE	18.88	0.13	1.02	1.22	1.00	0.07	0.00	1.00	1.00	1.00	1.00
2	BE	11.76	0.11	0.83	1.07	2.01	0.14	0.10	.99	.99	1.00	1.00
3	BE + CE	7.12	0.09	0.43	1.14	3.43	0.49	0.67	.76	.64	1.00	1.00
4	BF + CE	6.65	0.08	0.52	1.25	3.90	0.50	1.08	.74	.65	.99	1.00
5	BE + BF	4.11	0.07	0.40	0.79	5.42	0.37	1.69	.92	.92	1.00	.96
6	CE + EF	2.86	0.06	0.34	0.51	6.63	0.59	1.52	.68	.66	1.00	.95
7	BE + BF + CE	2.55	0.05	0.24	0.62	8.58	0.65	7.37	.70	.65	1.00	.93
8	CE + ADE	1.86	0.05	0.25	0.30	8.71	0.83	1.54	.57	.55	.95	.96
9	BE + EF	1.73	0.04	0.25	0.37	9.51	0.94	3.37	.53	.52	.93	.93
10	BE + ADE	1.14	0.04	0.18	0.22	12.26	1.40	3.19	.39	.24	.55	.29

Note: Posterior model probabilities $\boldsymbol{\pi}$ are shown in percent. Mean($\hat{\boldsymbol{\pi}}$) and SD($\hat{\boldsymbol{\pi}}$) were computed across 100 replications. All models include the six main effects, A: smoking, B: strenuous mental work, C: strenuous physical work, D: systolic blood pressure, E: ratio of α and β lipoproteins, F: family anamnesis of coronary heart disease, and the first-order interactions AC, AD, AE, BC, and DE.

6. CONCLUSION

We proposed a novel approach for estimating the precision of transdimensional MCMC output. Essentially, a Markov model is fitted to the observed model-indicator variable $z^{(t)}$ to obtain posterior samples of the corresponding stationary distribution. We showed that the method corrects for autocorrelation in a given sequence $z^{(t)}$ and provides a good assessment of estimation uncertainty. Importantly, the method does not require output of multiple independent MCMC chains and thus reduces the computational costs for adaption and burn-in. Besides being useful for transdimensional MCMC output, the method provides an estimate of the precision and effective sample size of discrete parameters in MCMC samplers in general. Thereby, researchers can easily decide whether the obtained precision is sufficiently high for substantive applications of interest.

ACKNOWLEDGMENTS

Daniel W. Heck was supported by the Grant Er 224/2-2 from the Deutsche Forschungsgemeinschaft (DFG) and the University of Mannheim’s Graduate School of Economic and Social Sciences, also funded by the DFG.

REFERENCES

- Anderson, T. W. and L. A. Goodman (1957). “Statistical Inference about Markov Chains.” In: *The Annals of Mathematical Statistics* 28, pp. 89–110.
- Arnold, R., Y. Hayakawa, and P. Yip (2010). “Capture–recapture Estimation Using Finite Mixtures of Arbitrary Dimension.” In: *Biometrics* 66, pp. 644–655.
- Barker, R. J. and W. A. Link (2013). “Bayesian Multimodel Inference by RJMCMC: A Gibbs Sampling Approach.” In: *The American Statistician* 67, pp. 150–156.

- Brooks, S. P. and P. Giudici (2000). “Markov Chain Monte Carlo Convergence Assessment via Two-Way Analysis of Variance.” In: *Journal of Computational and Graphical Statistics* 9, pp. 266–285.
- Brooks, S. P., P. Giudici, and G. O. Roberts (2003). “Efficient Construction of Reversible Jump Markov Chain Monte Carlo Proposal Distributions.” In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65, pp. 3–39.
- Brooks, S., P. Giudici, and A. Philippe (2003). “Nonparametric Convergence Assessment for MCMC Model Selection.” In: *Journal of Computational and Graphical Statistics* 12, pp. 1–22.
- Burke, C. J. and M. Rosenblatt (1958). “A Markovian Function of a Markov Chain.” In: *The Annals of Mathematical Statistics* 29, pp. 1112–1122.
- Carlin, B. P. and S. Chib (1995). “Bayesian Model Choice via Markov Chain Monte Carlo Methods.” In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57, pp. 473–484.
- Castelloe, J. M. and D. L. Zimmerman (2002). “Convergence Assessment for Reversible Jump MCMC Samplers.” Department of Statistics and Actuarial Science, University of Iowa.
- Cowles, M. K. and B. P. Carlin (1996). “Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review.” In: *Journal of the American Statistical Association* 91, pp. 883–904.
- Dellaportas, P., J. J. Forster, and I. Ntzoufras (2000). “Bayesian Variable Selection Using the Gibbs Sampler.” In: *Generalized Linear Models: A Bayesian Perspective*. Ed. by D. K. Dey, S. K. Ghosh, and B. K. Mallick. New York: Marcel Dekker, Inc., pp. 273–286.
- Dellaportas, P. and J. J. Forster (1999). “Markov Chain Monte Carlo Model Determination for Hierarchical and Graphical Log-Linear Models.” In: *Biometrika* 86, pp. 615–633.

- Dellaportas, P., J. J. Forster, and I. Ntzoufras (2002). “On Bayesian Model and Variable Selection Using MCMC.” In: *Statistics and Computing* 12, pp. 27–36.
- Eddelbüttel, D. and C. Sanderson (2014). “RcppArmadillo: Accelerating R with High-Performance C++ Linear Algebra.” In: *Computational Statistics and Data Analysis* 71, pp. 1054–1063.
- Edwards, D. and T. Havránek (1985). “A Fast Procedure for Model Search in Multidimensional Contingency Tables.” In: *Biometrika* 72, pp. 339–351.
- Forster, J. J., R. C. Gill, and A. M. Overstall (2012). “Reversible Jump Methods for Generalised Linear Models and Generalised Linear Mixed Models.” In: *Statistics and Computing* 22, pp. 107–120.
- Frühwirth-Schnatter, S. (2001). “Markov Chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models.” In: *Journal of the American Statistical Association* 96, pp. 194–209.
- Gelman, A. and D. B. Rubin (1992). “Inference from Iterative Simulation Using Multiple Sequences.” In: *Statistical Science* 7, pp. 457–472.
- Gong, L. and J. M. Flegal (2016). “A Practical Sequential Stopping Rule for High-Dimensional Markov Chain Monte Carlo.” In: *Journal of Computational and Graphical Statistics* 25, pp. 684–700.
- Green, P. J. (1995). “Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination.” In: *Biometrika* 82, pp. 711–732.
- Healy, M. J. R. (1988). *GLIM: An Introduction*. UK: Clarendon Press.
- Heck, D. W. et al. (2017). *MCMCprecision: Precision of Discrete Variables in Transdimensional MCMC*. <https://github.com/danheck/MCMCprecision>.
- Heidelberger, P. and P. D. Welch (1981). “A Spectral Method for Confidence Interval Generation and Run Length Control in Simulations.” In: *Communications of the ACM* 24, pp. 233–245.
- Jeffreys, H. (1961). *Theory of Probability*. New York: Oxford University Press.

- Karnesis, N. (2014). “Bayesian Model Selection for LISA Pathfinder.” In: *Physical Review D* 89.
- Kass, R. E. and A. E. Raftery (1995). “Bayes Factors.” In: *Journal of the American Statistical Association* 90, pp. 773–795.
- Kuo, L. and B. Mallick (1998). “Variable Selection for Regression Models.” In: *Sankhyā: The Indian Journal of Statistics, Series B* 60, pp. 65–81.
- Lodewyckx, T. et al. (2011). “A Tutorial on Bayes Factor Estimation with the Product Space Method.” In: *Journal of Mathematical Psychology* 55, pp. 331–347.
- Lopes, H. F. and M. West (2004). “Bayesian Model Assessment in Factor Analysis.” In: *Statistica Sinica* 14, pp. 41–67.
- Minka, T. (2000). *Estimating a Dirichlet Distribution*. Technical Report.
- Ntzoufras, I. (2002). “Gibbs Variable Selection Using BUGS.” In: *Journal of Statistical Software* 7, pp. 1–19.
- Ntzoufras, I., P. Dellaportas, and J. J. Forster (2003). “Bayesian Variable and Link Determination for Generalised Linear Models.” In: *Journal of Statistical Planning and Inference*. Special issue I: Model Selection, Model Diagnostics, Empirical Bayes and Hierarchical Bayes 111, pp. 165–180.
- Opgen-Rhein, R., L. Fahrmeir, and K. Strimmer (2005). “Inference of Demographic History from Genealogical Trees Using Reversible Jump Markov Chain Monte Carlo.” In: *BMC Evolutionary Biology* 5, p. 6.
- Overstall, A. M. and R. King (2014a). “A Default Prior Distribution for Contingency Tables with Dependent Factor Levels.” In: *Statistical Methodology* 16, pp. 90–99.
- Overstall, A. and R. King (2014b). “Conting: An R Package for Bayesian Analysis of Complete and Incomplete Contingency Tables.” In: *Journal of Statistical Software* 58, pp. 1–27.

- Plummer, M. (2003). “JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling.” In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Vol. 124. Vienna, Austria, p. 125.
- Plummer, M. et al. (2006). “CODA: Convergence Diagnosis and Output Analysis for MCMC.” In: *R News* 6, pp. 7–11.
- Sanderson, C. and R. Curtin (2016). “Armadillo: A Template-Based C++ Library for Linear Algebra.” In: *Journal of Open Source Software* 1, p. 26.
- Scott, J. G. and J. O. Berger (2010). “Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem.” In: *The Annals of Statistics* 38, pp. 2587–2619.
- Sisson, S. A. and Y. Fan (2007). “A Distance-Based Diagnostic for Trans-Dimensional Markov Chains.” In: *Statistics and Computing* 17, pp. 357–367.
- Sisson, S. A. (2005). “Transdimensional Markov Chains.” In: *Journal of the American Statistical Association* 100, pp. 1077–1089.
- Stephens, M. (2000). “Bayesian Analysis of Mixture Models with an Unknown Number of Components- an Alternative to Reversible Jump Methods.” In: *The Annals of Statistics* 28, pp. 40–74.
- Stewart, W. J. (2009). *Probability, Markov Chains, Queues, and Simulation*. Princeton, NJ: Princeton University Press.